限定领域语言模型训练语料的词类扩展方法①

黄韵竹，韦 玮，罗杨宇，李成荣

(中国科学院 自动化研究所，北京 100190)

摘 要：限定领域的语言模型训练语料的搜集需要耗费大量的人力物力，如果语料搜集不充分，往往会造成数据稀疏的问题。解决该问题的方法有两种：1、采用数据平滑算法，降低模型的困惑度；2、对训练语料进行扩展。探索了对语言模型的训练语料进行半自动扩展的方法。该方法通过计算互信息将非限定领域的大规模语料分成若干词类，生成大词类表；再将该表中领域相关的词类提取出来，进行手动删减之后用于对限定领域的语言模型进行参数估计。实验表明，将该方法用于语音识别系统，能有效缩短语言模型训练语料的搜集时间，提高系统的识别率。

关键词：语料扩展；互信息；语言模型；语音识别；词类

Word-Class Expansion Method About Training Corpus of Language Modal in Restrcited Domain

HUANG Yun-Zhu, WEI Wei, LUO Yang-Yu, LI Cheng-Rong

(Institute of Autumation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: It is time-consuming to collect training corpus of language modal in restricted domain. The insufficiency of corpus will lead to the problem of training data sparsity. There are two common methods to solve this problem. One is reducing the complexion of modal through data smoothing. The other is expanding the corpus. In this paper, a semiautomatic method to expand training corpus of language modal is proposed. A large list of word classes is generated by calculating the mutual information of non-restricted areas corpus in large scale. Then, those word classes related to the restricted domain is extracted and manually cut out to estimate parameters of language modal. Experimental results show that the method could effectively solve the problem of training data sparsity, and improve the recognition rate of speech recognition system.

Key words: corpus expansion; mutual information; language modal; speech recognition; word classes

1 引言

分词后的文本语料可以用来对语音识别系统的语言模型进行参数估计，被称为语言模型的训练语料。在限定领域的语音识别系统中，按照语言模型训练语料的来源，可将其分成实际场景语料和非实际场景语料两大类。实际场景语料通过对实际的应用场景录音并整理来获得，非实际场景语料的获得渠道多种多样，比如可以采用征集语料的方式，FSN 方式[1]等等。

语料搜集不充分导致的训练数据稀疏成为限定领域语言模型的主要难点之一。针对这一问题的研究主要集中在两个方面：一方面是采用数据平滑算法，另一方面是对训练语料进行扩展。数据平滑算法不能从根源解决数据稀疏问题，只能在一定程度上缓解数据稀疏带来的问题，而且有的算法本身存在一些缺点[1]。因此，扩展语料以提高语言模型性能的研究值得重视。

本文提出了一种限定领域语言模型语料扩展方法，用这种方法能够大大扩展语言模型的训练语料，实验证明，该方法能有效的提高系统的识别率，并且

① 收稿时间:2011-03-09;收到修改稿时间:2011-03-30

对原始语料集内词的识别率影响不大。这里的原始语料是指采用特定的扩展方法扩展之前已有的限定领域语料。

本文的内容安排如下：第 2 节讨论限定领域语言模型的训练语料搜集需要注意的一些问题。第 3 节介绍同义词语言模型语料扩展方法和基于词类的语料扩展的方法，分析了词类扩展方法的优越性。第 4 节用实验说明基于词类的语料扩展方法对语音识别系统的影响。最后是结束语。

2　扩展要求

我们认为，对限定领域而言，扩展语言模型的训练语料应当满足一定的要求:

第一，扩展出的语料应该尽量保持原始语料的语言风格。这里的语言风格主要指文体，用词等方面。

第二，加入扩展语料后训练生成的语言模型，应充分保证原始系统集内词的识别率。

以上两点要求可以作为选取语料扩展方法的一个指导思想。

3　语料扩展方法

一般而言，语言模型的训练语料是经过分词处理的句子，例如：“你/会/数数/吗”。语料扩展可以分为词级的扩展和句子级的扩展，本文提出的语料扩展方法都是词级的。针对词级扩展的问题，我们尝试了两种方法：(1)利用同义词词林[2]进行同义词扩展；(2)利用词分类算法，对已有的大规模文本进行词分类，对分类结果提取出领域相关的词类，达到扩展语料的目的。本文重点介绍语言模型训练语料的词类扩展方法，结合第 2 节提出的扩展要求，通过与同义词扩展方法作比较，分析出了词类扩展方法的优越性。

3.1 同义词扩展

该方法利用了同义词词林，对原始语料词典中的所有词查找同义词，从而达到语料扩展的目的。

然而，该方法具有一定的局限性。首先，扩展出的语料不能包含新的意思，只是原始语料的不同表达。其次，由于同义词词林本身的局限性，扩展出的词很多不符合现代用语习惯。

例如，原始的语料为“今天天气不错”，分词后为“今天/天气/不错”。进行同义词扩展的结果见表 1。

其中同义词词林 1 是原始的同义词词林；同义词词林 2 是我们对原始词林进行裁减，去掉了一些冷僻词之后得到的词林。

由表 1 中的实验结果分析可以得出，扩展出的同义词过多，很多同义词不能够在句子中取代原始词。如果要将扩展出的同义词用于语言模型的训练，还需要进一步想办法对同义词词林和扩展后的结果进行大量的删减。

表 1　同义词扩展结果

同义词词林	原始词	同义词个数	同义词举例
1	今天	35	本、当前、即
	天气	5	天候、天道、天
	不错	82	帅、地道、白璧无瑕
2	今天	16	当今、当前、而今
	天气	2	气候、气象
	不错	44	精炼、精良、精美

3.2 词类扩展

词类扩展是本文提出的一种语料扩展方法。它的基本思想是：首先采用基于互信息的自下而上合并的词分类方法，对通用的大规模语料库（如 863 大规模语料库等）进行词分类，形成大词类表。然后根据这个大词类表查找原始语料中的词的词类，生成小词类表，从而达到语料扩展的目的。

3.2.1 互信息的计算公式

在自然语言中，词类{$C_1, C_2, C_3, \cdots, C_n$}的分布满足随机分布，词类的互信息计算公式如下[3,4]：

$$Ma(f)=\sum_{Ci}\sum_{Cj}P(Ci,Cj)\times\log\frac{P(Ci,Cj)}{P(Ci)P(Cj)} \quad (1)$$

其中 $P(Ci),P(Cj)$ 及 $P(Ci,Cj)$ 分别表示词类 Ci，Cj 出现的概率及它们同现的概率。具体计算方法如下：

$$P(Ci)=\frac{\sum_{w\in Ci}N_w}{N_{total}} \quad (2)$$

$$P(Ci,Cj)=\frac{\sum_{w1\in Ci}\sum_{w2\in Cj}N(w_1,w_2)}{N_{bitotal}} \quad (3)$$

其中 N_w 表示词 w 在语料库中出现的次数，$N(w_1,w_2)$ 表示词对 w_1,w_2 在语料库中按顺序出现的次数，N_{total} 表示语料库中包含所有词的个数（含重复），$N_{bitotal}$ 表示词对的总数（含重复）。

3.2.2 分类算法的实现

对于大词汇量的语料而言，必须考虑的一个重要

因素就是算法的时间复杂度。本文采用自下而上合并的分类算法，算法的流程如下：

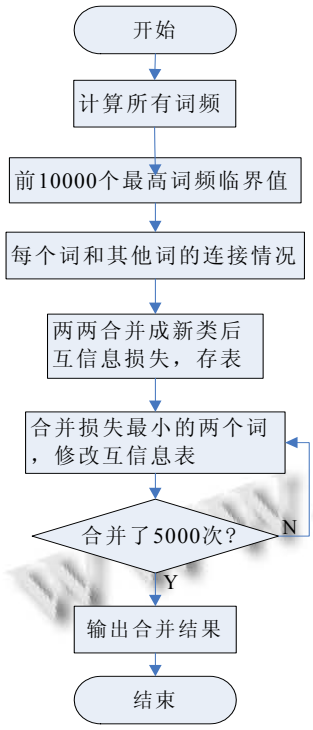

图1 词分类算法流程

1) 计算所有词出现的频率。

2) 计算前 10000 个（经验值）最高词频的临界值，即把所有词的出现次数按由高到低排列，排到第 10000 的那个次数。

3) 计算每个词和其他词的连接情况（分为左连接和右连接）。

4) 将每个词看成单独的词类，计算前 10000 个最高词频的词两两合并成词类后互信息的损失，生成互信息减少量二维表（以下简称互信息表）。

5) 将互信息损失最少的词归成一类，修改互信息表。

6) 重复步骤 5)，共 5000 次（经验值），生成最后的词类表。

本算法最后生成了词类 2122 类（只计算两个词以上的词类）。生成的词类如下：

Class1：除此之外 尽管如此 相比之下 遗憾的是 无论如何

Class76：崇高 高尚 神圣 永恒 纯洁 天真 活泼 善良

Class112：安徽省 湖南省 甘肃省 云南省 福建省 四川省 湖北省 陕西省

Class212：扩散 蔓延 扩展 延伸

3.3 两种方法的对比分析

以上分别介绍了同义词扩展方法和词类扩展方法，由扩展结果可以看出，词类扩展方法具有明显的优越性。第一，词类扩展方法扩展出的词语，更好的满足了第 2 节提出的第 1 点扩展要求。第二，词类扩展方法扩展出的词，不仅包含同义词，还包含了更多的“同类词”。

3.4 基于词类的语言模型

词类扩展方法是对词进行扩展，刚好可以用在基于词类的语言模型上，达到语料扩展的目的。

基于词类的语言模型是对基于词的语言模型的改进。假设属于类，由语言模型计算 n 元概率 $P(w_i|w_{i-n+1}^{i-1})$ 公式如下[5]：

$$P(w_i|w_{i-n+1}^{i-1}) = P(w_i|C_j)P(C_j|C_{j-n+1}^{j-1}) \quad (4)$$

一般情况下，由于词类的数目小于词的数目，这样，在估计 n 元概率时面临的数据稀疏问题在一定程度得到缓解，提高了对训练语料中未出现的词串的预测能力。此外，还压缩了语言模型的尺寸。

如果把每个词看成一个类就回退到基于词的语言模型了，因此基于词的语言模型可以看成是基于词类的语言模型的一个特例。

4 系统实现及实验结果分析

4.1 统实现

词类扩展利用了 863 大规模语料库，分词后大小为 1.39G，共有 40089 个词（不含重复）。原始语料的话题限定为日常对话聊天。

第一步，首先对 863 大规模语料库进行分词处理。然后对分词后的结果按照第 3 节中的算法进行词分类生成大词类表。再在大词类表中查找待扩展的原始语料中的词的词类信息，生成小词类表 1。最后进行手动删减，完成语料扩展。

第二步，对原始语料进行词分类，生成小词类表 2。

第三步，合并小词类表 1 和表 2。对扩展后的语料进行训练，生成基于词类的语言模型。

4.2 实验及结果分析

实验通过对比词类扩展前（以下称为原始系统）和词类扩展后（以下称为新系统）各自生成的语言模

型对语音识别系统的性能指标，验证了基于词类扩展的语言模型对系统的影响。话题限定为日常对话聊天，测试语音共 10 人录音，5 女 5 男，每人 200 句，共 2000 句。采样率为 16kHz，声道为单声道，采样精度为 16bit。

实验 1 主要考察新系统与原始系统相比，识别率的变化情况[6]。识别结果如表 2 所示：

表 2　原始系统和新系统各项指标对比

	原始系统	新系统	变化量
Mcr (字匹配率)	65.57%	66.61%	1.04%
WER (字错误率)	35.13%	34.09%	-1.04%
Ser (替换错误率)	27.72%	26.82%	-0.90%
Ier (插入错误率)	0.70%	0.70%	0.00%
Der (删除错误率)	6.71%	6.56%	-0.15%

由表 2 可以看出，新系统的字匹配率比原始系统有所提高，各项错误指标比原始系统有所降低，总的性能得到一定改善。

实验 2 主要考察新系统对集内词的识别率的影响，以及对原始系统不能识别的词的识别情况。

$$p(Nr \mid Or) = \frac{N_{Nr|Or}}{N_{Or}} \tag{5}$$

$$p(Nr \mid Ow) = \frac{N_{Nr|Ow}}{N_{Ow}} \tag{6}$$

其中，N_{Or} 表示原始系统识别正确的字数，N_{Ow} 表示原始系统识别错误的字数。$N_{Nr \mid Or}$ 表示原始系统识别正确的情况下，新系统也识别正确的字数。$N_{Nr \mid Ow}$ 表示原始系统识别错误的情况下，新系统识别正确的字数。

表 3　新系统对原始系统的影响

$p(Nr \mid Or)$	98.38%
$p(Nr \mid Ow)$	10.51%

由表 3 中的实验结果可以看出，新系统对原始系统集内词的识别率影响不大，验证了本文第 2 节提出的第二点要求。此外，新系统还识别出了很多原始系统不能识别出的字。

该实验的限定话题是日常对话聊天，实验取得了较好的结果。本方法对其他限定领域也同样适用。对一些实际应用中同类词出现较多，而句式又相对较少的领域，比如介绍各地地理状况和风土人情的问答系统，应该能够有更好的实验结果。

5　结论

本文提出了一种语言模型语料的半自动扩展方法，采用该方法可以有效的扩展语料中词的同类词（即采用本论文提出的词类生成方法生成的同一个词类中的词），从而达到扩展语料的目的。实验结果表明，该方法可以提高语音识别系统识别率，同时又对原始系统集内词的识别率影响不大。

为了提高语音识别系统整体的识别率，进一步降低新系统对原始系统集内词识别率的影响，下一步工作还可以从以下几方面入手：

1. 对基于词类的语言模型进行改进，降低采用词类扩展生成的新词的权重。

2. 适当的选择生成大词类表用到的大规模语料，使得大规模语料和该限定领域风格更加相似。

3. 对生成大词类表时算法中的循环次数的经验值进行调整（本文取 5000 次），使得结果更加符合实际需要。

值得指出的是，该方法不仅能应用于语音识别系统，对光学字符识别（OCR），汉字智能输入等领域的语言模型语料扩展也提供了一种可能的方法。

参考文献

1 熊军军,李成荣.实际场景语料和 FSN 语料的平衡方法.清华大学学报(自然科学版),2008,48(1):730−734.

2 梅家驹,等.同义词词林.上海:上海辞书出版社出版, 1983.

3 Brown PF, deSouza PV, Mercer RL, et al. Class-Based n-gram Models of Natural Language. Computational Linguistics, 1992,18(4):467−479.

4 赵石顽,夏莹,马少平,王昱,苏中.基于统计的中文词分类.毛剑琴.第三届全球智能控制与自动化大会论文集.合肥:中国科学技术大学出版社,2000:2753−2756.

5 陈振标,徐波.限定领域的语言模型.第七届全国人机语音通讯学术会议.厦门,2003.27−30.

6 徐波,孙甲松,李爱军,徐明星,黄泰翼,鲍怀翘,尹波,吴志刚.中文语音识别系统通用技术规范.中华人民共和国国家标准.北京:中国标准出版社,2007.